The Thesis Committee for Pandian Raju
Certifies that this is the approved version of the following Thesis:

PebblesDB: Building Key-Value Stores using Fragmented Log-Structured Merge Trees

APPROVED BY

SUPERVISING COMMITTEE:

Vijay Chidambaram Velayudhan Pillai, Supervisor

Simon Peter

PebblesDB: Building Key-Value Stores using Fragmented Log-Structured Merge Trees

by

Pandian Raju

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2018

Dedicated to my parents, Mr. Raju Sankaran and Mrs. Shanthi Pandian,
and my brother, Ravi Sankar Raju.

Acknowledgments

I would first like to thank my thesis advisor Prof. Vijay Chidambaram of the department of Computer Science at The University of Texas at Austin, without whom this work wouldn't have been possible. He was more like a mentor than an advisor, giving me complete independence to work on the research, and at the same time guiding me whenever needed. He was available every time I had some question or was stuck at some point in research. I had the pleasure of working under his advising, and would like to thank him again for the multitude of things that I learnt from him during my Masters.

I would like to thank Prof. Simon Peter of the department of Computer Science at The University of Texas at Austin, for serving as the second reader and for giving feedback on this thesis. I would also like to thank the members of the LASR (Laboratory for Advanced Systems Research) group and the Systems and Storage Lab (SASLab) at The University of Texas at Austin for their feedback and guidance on this work.

I would also like to thank the program committee members of the Symposium on Operating Systems Principles (SOSP'17) conference, who reviewed the paper (on PebblesDB) and provided valuable and constructive feedback, which made the work on this thesis stronger. Particularly, I would like to thank Dr. Frans Kaashoek, who served as the shepherd for our paper and

helped us in refining the work in the paper.

I would like to thank my other co-authors of the paper, Rohan Kadekodi, Masters student at The University of Texas at Austin, for helping to get more evaluation numbers, and Ittai Abraham, researcher in VMware Research Group, for his valuable insights in the high level design of the work in this thesis. I would also like to thank VMware and Facebook for their generous donations supporting this work.

I am fortunate to have been blessed with a wonderful family, who believes in whatever I pursue. I would like to thank my brother Ravi Sankar, a role model in my life, who inspired me to pursue my Masters degree. I would also like to thank a multitude of friends here at The University of Texas at Austin, who gave me moral support all throughout my Masters course.

Finally, I would like to thank God for giving me an opportunity to pursue Masters at The University of Texas at Austin, to work with Prof. Vijay Chidambaram, to work on the exciting work in this thesis, and for blessing me with a wonderful life.

PebblesDB: Building Key-Value Stores using Fragmented Log-Structured Merge Trees

Pandian Raju, M.S.Comp.Sci
The University of Texas at Austin, 2018

Supervisor: Vijay Chidambaram Velayudhan Pillai

Key-value stores such as LevelDB and RocksDB offer excellent write throughput, but suffer high write amplification. The write amplification problem is due to the Log-Structured Merge Trees data structure that underlies these key-value stores. To remedy this problem, this thesis presents a novel data structure that is inspired by Skip Lists, termed Fragmented Log-Structured Merge Trees (FLSM). FLSM introduces the notion of guards to organize logs (sstables or files containing the data on storage), and avoids rewriting data in the same level. Theoretically, we show how FLSM can address the problem of write amplification.

We build PebblesDB, a high-performance key-value store, by modifying HyperLevelDB to use the FLSM data structure. We evaluate PebblesDB using micro-benchmarks and show that for write-intensive workloads, PebblesDB reduces write amplification by 2.4-3× compared to RocksDB, while increasing write throughput by 6.7×. We evaluate PebblesDB extensively under a variety of benchmarks, workload patterns, and environmental factors and analyze

how it performs in different scenarios. We modify two widely-used NoSQL stores, MongoDB and HyperDex, to use PebblesDB as their underlying storage engine. Evaluating these applications using the YCSB benchmark shows that throughput is increased by 18-105% when using PebblesDB (compared to their default storage engines) while write IO is decreased by 35-55%.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Key-value stores have become a fundamental part of the infrastructure for modern systems. Much like how file systems are an integral part of operating systems, distributed systems today depend on key-value stores for storage. For example, key-value stores are used to store state in graph databases [31, 21], task queues [5, 55], stream processing engines [7, 51], application data caching [43, 35], event tracking systems [46], NoSQL stores [40, 18], and distributed databases [30]. Improving the performance of key-value stores has the potential to impact a large number of widely-used data intensive services.

Great progress has been made in improving different aspects of key-value stores such as memory efficiency [59, 9, 42, 34, 17] and energy efficiency [6]. One fundamental problem that remains is the high write amplification of key-value stores for write-intensive workloads. Write amplification is the ratio of total write IO performed by the store to the total user data. High write amplification increases the load on storage devices such as SSDs, which have limited write cycles before the bit error rate becomes unacceptable [3, 26, 39]. With the increasing size of user data sets (*e.g,.* Pin-

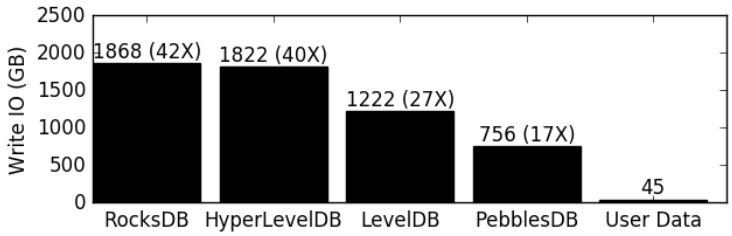


Figure 1.1: **Write Amplification**. The figure shows the total write IO (in GB) for different key-value stores when 500 million key-value pairs (totaling 45 GB) are inserted or updated. The write amplification is indicated in parenthesis.

terest's stateful systems process tens of petabytes of data every day [46]), high write amplification results in frequent device wear out and high storage costs [41]. Write amplification also reduces write throughput: in the RocksDB [20] key-value store, it results in write throughput being reduced to 10% of read throughput [54]. Thus, reducing write amplification will both lower storage costs and increase write throughput.

Figure 1.1 shows the high write amplification (ratio of total IO to total user data written) that occurs in several widely-used key-value stores when 500 million key-value pairs are inserted or updated in random order. Techniques from prior research tackling write amplification have not been widely adopted since they either require specialized hardware [38, 56] or sacrifice other aspects such as search (range query) performance [58]. Conventional wisdom is that reducing write amplification requires sacrificing either write or read throughput [34]. In today's low-latency, write-intensive environments [27], users are not willing to sacrifice either.

Key-value stores such as LevelDB [25] and RocksDB are built on top of the log-structured merge trees [44] (LSM) data structure, and their high write amplification can be traced back to the data structure itself (chapter 2). LSM stores maintain data in sorted order on storage, enabling efficient querying of data. However, when new data is inserted into an LSM-store, existing data is rewritten to maintain the sorted order, resulting in large amounts of write IO.

This thesis presents a novel data structure, termed the *Fragmented Log-Structured Merge Trees* (FLSM), which combines ideas from the Skip List [48, 47] and Log-Structured Merge trees data structures along with a novel compaction algorithm. FLSM strikes at the root of write amplification by drastically reducing (and in many cases, eliminating) data rewrites, instead *fragmenting* data into smaller chunks that are organized using *guards* on storage (chapter 3). Guards allow FLSM to find keys efficiently. Write operations on LSM stores are often stalled or blocked while data is compacted (rewritten for better read performance); by drastically reducing write IO, FLSM makes compaction significantly faster, thereby increasing write throughput.

Building a high-performance key-value store on top of the FLSM data structure is not without challenges; the design of FLSM trades read performance for write throughput. This thesis presents PebblesDb [49], a modification of the HyperLevelDB [29] key-value store that achieves the trifecta of low write amplification, high write throughput, and high read throughput. PebblesDb employs a collection of techniques such as parallel seeks, aggressive seek-based compaction, and sstable-level bloom filters to reduce the overheads

inherent to the FLSM data structure (chapter 4). Although many of the techniques PebblesDb employs are well-known, together with the FLSM data structure, they allow PebblesDb to achieve excellent performance on both read-dominated and write-dominated workloads.

PebblesDb outperforms mature, carefully engineered key-value stores such as RocksDB and LevelDB on several workloads (chapter 5). On the `db_bench` micro-benchmarks, PebblesDb obtains 6.7× the write throughput of RocksDB and 27% higher read throughput, while doing 2.4-3× less write IO. When the NoSQL store MongoDB [40] is configured to use PebblesDb instead of RocksDB as its storage engine, MongoDB obtains the same overall performance on YCSB benchmark [16] while doing 37% lesser IO (chapter 5).

While the FLSM data structure is useful in many scenarios, it is not without its limitations. On a fully compacted key-value store, PebblesDb incurs a 30% overhead for small range queries. While the overhead drops to 11% for large range queries, the FLSM data structure is not the best fit for workloads which involve a lot of range queries after an initial burst of writes. Note that PebblesDb does not incur an overhead if the range queries are interspersed with writes.

The work in thesis was published in SOSP 2017 [49], done in collaboration with Rohan Kadekodi (Masters student at The University of Texas at Austin), Prof. Vijay Chidambaram (Assistant Professor at The University of Texas at Austin), and Ittai Abraham (researcher in VMware Research).

Chapter 2

Background

This chapter provides some background on key-values stores and log-structured merge trees. It first describes common operations on key-values stores (section 2.1) and discusses why log-structured merge trees are used to implement key-value stores in write-intensive environments (section 2.2). It shows that the log-structured merge tree data structure fundamentally leads to large write amplification.

2.1 Key-Value Store Operations

Get. The `get(key)` operation returns the latest value associated with `key`.

Put. The `put(key, value)` operation stores the mapping from `key` to `value` in the store. If `key` was already present in the store, its associated value is updated.

Iterators. Some key-value stores such as LevelDB provide an iterator over the entire key-value store. `it.seek(key)` positions the iterator `it` at the smallest key $\geq$ `key`. The `it.next()` call moves `it` to the next key in sequence. The `it.value()` call returns the value associated with the key at the current

iterator position. Most key-value stores allow the user to provide a function for ordering keys.

Range Query. The `range_query(key1, key2)` operation returns all key-value pairs falling within the given range. Range queries are often implemented by doing a `seek()` to `key1` and doing `next()` calls until the iterator passes `key2`.

2.2 Log-Structured Merge Trees

Embedded databases such as KyotoCabinet [32] and BerkeleyDB [45] are typically implemented using B+ Trees [14]. However, B+ Trees are a poor fit for write-intensive workloads since updating the tree requires multiple random writes (10-100× slower than sequential writes). Inserting 100 million key-value pairs into KyotoCabinet writes 829 GB to storage (61× write amplification). Due to the low write throughput and high write amplification of B+ Trees, developers turned to other data structures for write-intensive workloads.

The log-structured merge trees (LSM) data structure [44] takes advantage of high sequential bandwidth by *only* writing sequentially to storage. Writes are batched together in memory and written to storage as a sequential log (termed an *sstable*). Each sstable contains a sorted sequence of keys.

Sstables on storage are organized as hierarchy of *levels*. Each level contains multiple sstables, and has a maximum size for its sstables. In a 5-

level LSM, Level 0 is the *lowest* level and Level 5 is the *highest* level. The amount of data (and the number of sstables) in each level increases as the levels get higher. The last level in an LSM may contain hundreds of gigabytes. Application data usually flows into the lower levels and is then compacted into the higher levels. The lower levels are usually cached in memory.

LSM maintains the following invariant at each level: all sstables contain disjoint sets of keys. For example, a level might contain three sstables: {1..6}[1], {8..12}, and {100..105}. Each key will be present in exactly one sstable on a given level. As a result, locating a key requires only two binary searches: one binary search on the starting keys of sstables (maintained separately) to locate the correct sstable and another binary search inside the sstable to find the key. If the search fails, the key is not present in that level.

LSM Operations. The `get()` operation returns the latest value of the key. Since the most recent data will be in lower levels, the key-value store searches for the key level by level, starting from Level 0; if it finds the key, it returns the value. Each key has a sequence number that indicates its version. Finding the key at each level requires reading and searching exactly one sstable.

The `seek()` and `next()` operations require positioning an iterator over the entire key-value store. This is implemented using multiple iterators (one per level); each iterator is first positioned inside the appropriate sstable in each level, and the iterator results are merged. The `seek()` operation requires

[1]Let {x..y} indicate a sstable with keys ranging from x to y

finding the appropriate sstables on each level, and positioning the sstable iterators. The results of the sstable iterators are merged (by identifying the smallest key) to position the key-value store iterator. The `next()` operation simply advances the correct sstable iterator, merges the iterators again, and re-positions the key-value store iterator.

The `put()` operation writes the key-value pair, along with a monotonically increasing sequence number, to an in-memory skip list [48] called the *memtable*. When the memtable reaches a certain size, it is written to storage as a sstable at Level 0. When a level obtains a threshold number of files, it is compacted into the next level. Assume Level 0 contains {2, 3} and {10, 12} sstables. If Level 1 contains {1, 4} and {9, 13} sstables, then during compaction, Level 1 sstables are rewritten as {1, 2, 3, 4} and {9, 10, 12, 13}, merging the sstables from Level 0 and Level 1. Compacting sstables reduces the total number of sstables in the key-value store and pushes colder data into higher levels. The lower levels are usually cached in memory, thus leading to faster reads of recent data.

Updating or deleting keys in LSM-based stores does not update the key in place, since all write IO is sequential. Instead, the key is inserted once again into the database with a higher sequence number; a delete key is inserted again with a special flag (often called a *tombstone* flag). Due to the higher sequence number, the latest version of the flag will be returned by the store to the user.

Write Amplification: Root Cause. Figure 2.1 illustrates compaction in

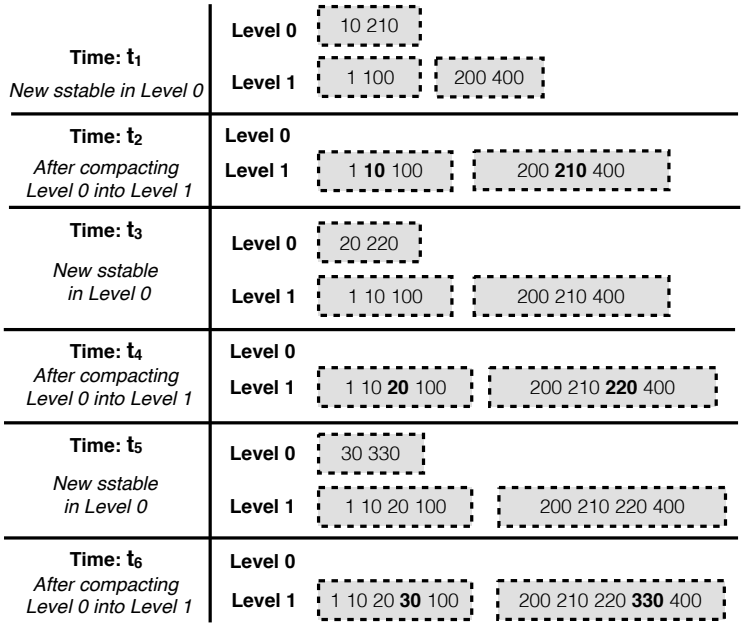


Figure 2.1: **LSM Compaction**. The figure shows sstables being inserted and compacted over time in a LSM.

a LSM key-value store. The key-value store contains two sstables in Level 1 initially. Let us assume that Level 0 is configured to hold only one sstable at a time; when this limit is reached, compaction is triggered. At time t_1, one sstable is added, and a compaction is triggered is at t_2. Similarly, sstables are added at t_3 and t_5 and compactions are triggered at t_4 and t_6. When compacting a sstable, all sstables in the next level whose key ranges *intersect* with the sstable being compacted are rewritten. In this example, since the key ranges of all Level 0 sstables intersect with key ranges of all Level 1 sstables, the Level 1 sstables are rewritten *every* time a Level 0 sstable is compacted. In this worst-case example, Level 1 sstables are rewritten **three** times while compacting a

single upper level. Thus, the high write amplification of LSM key-value stores can be traced to **multiple rewrites** of sstables during compaction.

The Challenge. A naive way to reduce write amplification in LSM is to simply not merge sstables during compaction but add new sstables to each level [19, 22]. However, read and range query performance will drop significantly due to two reasons. First, without merge, the key-value store will end up containing large number of sstables. Second, as multiple sstables can now contain the same key and can have overlapping key ranges in the same level, read operations will have to examine multiple sstables (since binary search to find the sstable is not possible), leading to large overhead.

Chapter 3

Fragmented Log-Structured Merge Trees

The challenge is to achieve three goals *simultaneously*: low write amplification, high write throughput, and good read performance. This chapter presents our novel data structure, Fragmented Log-structured Merge Trees (FLSM), and describes how it tackles this challenge.

FLSM can be seen as a blend of an LSM data structure with a Skip List along with a novel compaction algorithm that overall reduces write amplification and increases write throughput. The fundamental problem with log-structured merge trees is that sstables are typically re-written multiple times as new data is compacted into them. FLSM counters this by *fragmenting* sstables into smaller units. Instead of rewriting the sstable, FLSM's compaction simply appends a new sstable fragment to the next level. Doing so ensures that data is written *exactly once* in most levels; a different compaction algorithm is used for the the last few highest levels. FLSM achieves this using a novel storage layout and organizing data using *guards* (section 3.1). This chapter describes how guards are selected (section 3.2), how guards are inserted and deleted (section 3.3), how FLSM operations are performed (section 3.4), how FLSM can be tuned for different performance/write-IO trade-offs

(section 3.5), and its limitations (section 3.6).

3.1 Guards

In the classical LSM, each level contains sstables with disjoint key ranges (*i.e.,* each key will be present in exactly one sstable). The chief insight in this work is that maintaining this invariant is the root cause of write amplification, as it forces data to be rewritten in the same level. The FLSM data structure discards this invariant: each level can contain multiple sstables with overlapping key ranges, so that a key may be present in multiple sstables. To quickly find keys in each level, FLSM organizes the sstables into guards (inspired from the Skip-List data structure [48, 47]).

Each level contains multiple guards. Guards divide the key space (for that level) into disjoint units. Each guard G_i has an associated key K_i, chosen from among keys inserted into the FLSM. Each level in the FLSM contains more guards than the level above it; the guards get progressively more fine-grained as the data gets pushed deeper and deeper into the FLSM. As in a skip-list, if a key is a guard at a given level i, it will be a guard for all levels $> i$.

Each guard has a set of associated sstables. Each sstable is sorted. If guard G_i is associated with key K_i and guard G_{i+1} with K_{i+1}, an sstable with keys in the range $[K_i, K_{i+1})$ will be attached to G_i. Sstables with keys smaller than the first guard are stored in a special *sentinel* guard in each level. The last guard G_n in the level stores all sstables with keys $\geq K_n$. Guards within a

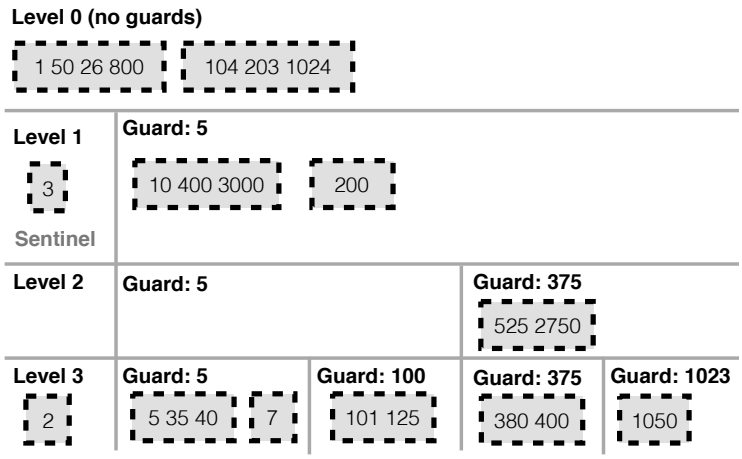


Figure 3.1: **FLSM Layout on Storage**. The figure illustrates FLSM's guards across different levels. Each box with dotted outline is an sstable, and the numbers represent keys.

level never have overlapping key ranges. Thus, to find a key in a given level, only one guard will have to be examined.

In FLSM compaction, the sstables of a given guard are (merge) sorted and then *fragmented* (partitioned), so that each child guard receives a new sstable that fits into the key range of that child guard in the next level.

Example. Figure 3.1 shows the state of the FLSM data structure after a few `put()` operations. We make several observations based on the figure:

- A `put()` results in keys being added to the in-memory memtable (not shown). Eventually, the memtable becomes full, and is written as an sstable to Level 0. Level 0 does not have guards, and collects together recently written sstables.

- The number of guards increases as the level number increases. The number of guards in each level does not necessarily increase exponentially.
- Each level has a sentinel guard that is responsible for sstables with keys $<$ than the first guard. In Figure 3.1, sstables with keys < 5 are attached to the sentinel guard.
- Data inside an FLSM level is *partially sorted*: guards do not have overlapping key ranges, but the sstables attached to each guard can have overlapping key ranges.

3.2 Selecting Guards

FLSM performance is significantly impacted by how guards are selected. In the worst case, if one guard contains all sstables, reading and searching such a large guard (and all its constituent sstables) would cause an un-acceptable increase in latency for reads and range queries. For this reason, guards are not selected statically; guards are selected probabilistically from inserted keys, preventing skew.

Guard Probability. When a key is inserted into FLSM, *guard probability* determines if it becomes a guard. Guard probability `gp(key,i)` is the probability that `key` becomes a guard at level `i`. For example, if the guard probability is 1/10, one in every 10 inserted keys will be randomly selected to be a guard. The guard probability is designed to be lowest at Level 1 (which has the fewest guards), and it increases with the level number (as higher levels

have more guards). Selecting guards in this manner distributes guards across the inserted keys in a smooth fashion that is likely to prevent skew.

Much like skip lists, if a key K is selected as a guard in level i, it becomes a guard for all higher levels $i+1, i+2$ etc. The guards in level $i+1$ are a strict superset of the guards in level i. Choosing guards in this manner allows the interval between each guard to be successively refined in each deeper level. For example, in Figure 3.1, key 5 is chosen as a guard for Level 1; therefore it is also a guard for levels 2 and 3.

FLSM selects guards out of inserted keys for simplicity; FLSM does not require that guards correspond to keys present in the key-value store.

Other schemes for selecting guards. The advantage of the current method for selecting guards is that it is simple, cheap to compute, and fairly distributes guards over inserted keys. However, it does not take into account the amount of IO that will result from partitioning sstables during compaction (this chapter will describe how compaction works shortly). FLSM could potentially select new guards for each level at compaction time such that sstable partitions are minimized; however, this could introduce skew. We leave exploring alternative selection schemes for future work.

3.3 Inserting and Deleting Guards

Guards are not inserted into FLSM synchronously when they are selected. Inserting a guard may require splitting an sstable or moving an sstable.

If a guard is inserted on multiple levels, work is generated on all those levels. For this reason, guards are inserted asynchronously into FLSM.

When guards are selected, they are added to an in-memory set termed the *uncommitted* guards. Sstables are not partitioned on storage based on (as of yet) uncommitted guards; as a result, FLSM reads are performed as if these guards did not exist. At the next compaction cycle, sstables are partitioned and compacted based on both old guards and uncommitted guards; any sstable that needs to be split due to an uncommitted guard is compacted to the next level. At the end of compaction, the uncommitted guards are persisted on storage and added to the full set of guards. Future reads will be performed based on the full set of guards.

We note that in many of the workloads that were tested, guard deletion was not required. A guard could become empty if all its keys are deleted, but empty guards do not cause noticeable performance degradation as `get()` and range query operations skip over empty guards. Nevertheless, deleting guards is useful in two scenarios: when the guard is empty or when data in the level is spread unevenly among guards. In the second case, consolidating data among fewer guards can improve performance.

Guard deletion is also performed asynchronously similar to guard insertion. Deleted guards are added to an in-memory set. At the next compaction cycle, sstables are re-arranged to account for the deleted guards. Deleting a guard G at level i is done lazily at compaction time. During compaction, guard G is deleted and sstables belonging to guard G will be partitioned and

appended to either the neighboring guards in the same level i or child guards in level $i+1$. Compaction from level i to $i+1$ proceeds as normal (since G is still a guard in level $i+1$). At the end of compaction, FLSM persists metadata indicating G has been deleted at level i. If required, the guard is deleted in other levels in a similar manner. Note that if a guard is deleted at level i, it should be deleted at all levels $< i$; FLSM can choose whether to delete the guard at higher levels $> i$.

3.4 FLSM Operations

Get Operations. A `get()` operation first checks the in-memory memtable. If the key is not found, the search continues level by level, starting with level 0. During the search, if the key is found, it is returned immediately; this is safe since updated keys will be in lower levels that are searched first. To check if a key is present in a given level, binary search is used to find the single guard that could contain the key. Once the guard is located, its sstables are searched for the key. Thus, in the worst case, a `get()` requires reading one guard from each level, and all the sstables of each guard.

Range Queries. Range queries require collecting all the keys in the given range. FLSM first identifies the guards at each level that intersect with the given range. Inside each guard, there may be multiple sstables that intersect with the given range; a binary search is performed on each sstable to identify the smallest key overall in the range. Identifying the next smallest key in the

range is similar to the merge procedure in merge sort; however, a full sort does not need to be performed. When the end of range query interval is reached, the operation is complete, and the result is returned to the user. Key-value stores such as LevelDB provide related operations such as `seek()` and `next()`; a `seek(key)` positions an iterator at the smallest key larger than or equal to `key`, while `next()` advances the iterator. In LSM stores, the database iterator is implemented via merging level iterators; in FLSM, the level iterators are themselves implemented by merging iterators on the sstables inside the guard of interest.

Put Operations. A `put()` operation adds data to an in-memory memtable. When the memtable gets full, it is written as a sorted sstable to Level 0. When each level reaches a certain size, it is compacted into the next level. In contrast to compaction in LSM stores, FLSM avoids sstable rewrites in most cases by partitioning sstables and attaching them to guards in the next level.

Key Updates and Deletions. Similar to LSM, updating or deleting a key involves inserting the key into the store with an updated sequence number or a deletion flag respectively. Reads and range queries will ignore keys with deletion flags. If the insertion of a key resulted in a guard being formed, the deletion of the key does not result in deletion of the related guard; deleting a guard will involve a significant amount of compaction work. Thus, empty guards are possible.

Compaction. When a guard accumulates a threshold number of sstables, it

is compacted into the next level. The sstables in the guard are first (merge) sorted and then partitioned into new sstables based on the guards of the next level; the new sstables are then attached to the correct guards. For example, assume a guard at Level 1 contains keys $\{1, 20, 45, 101, 245\}$. If the next level has guards 1, 40, and 200, the sstable will be partitioned into three sstables containing $\{1, 20\}$, $\{45, 101\}$, and $\{245\}$ and attached to guards 1, 40, and 200 respectively.

Note that in most cases, FLSM compaction does not rewrite sstables. This is the main insight behind how FLSM reduces write amplification. New sstables are simply added to the correct guard in the next level. There are two exceptions to the no-rewrite rule. First, at the highest level (*e.g,.* Level 5) of FLSM, the sstables have to be rewritten during compaction; there is no higher level for the sstables to be partitioned and attached to. Second, for the second-highest level (*e.g,.* Level 4), FLSM will rewrite an sstable into the same level if the alternative is to merge into a large sstable in the highest level (since we cannot attach new sstables in the last level if the guard is full). The exact heuristic is rewrite in second-highest-level if merge causes 25× more IO.

FLSM compaction is trivially parallelizable because compacting a guard only involves its descendants in the next level; the way guards are chosen in FLSM guarantees that compacting one guard never interferes with compacting another guard in the same level. For example, in Figure 3.1 if guard 375 in Level 2 is split into guards 375 and 1023 in Level 3, only these three guards are affected. Compacting guard 5 (if it had data) will not affect the on-going

compaction of guard 375 in any way. Thus, the compaction process can be carried out in parallel for different guard files at the same time. Parallel IO from compaction can be efficiently handled by devices such as flash SSDs that offer high random write throughput with multiple flash channels. Such parallel compaction can reduce the total time taken to compact significantly. A compacted key-value store has lower latency for reads; since parallel compaction gets the store to this state faster, it also increases read throughput.

3.5 Tuning FLSM

FLSM performance for reads and range queries depends upon a single parameter: the number of sstables inside each guard. If guards contain a large number of sstables, read and range query latencies become high. Therefore, FLSM provide users a knob to tune behavior, `max_sstables_per_guard`, which caps the maximum number of sstables present inside each guard in FLSM. When any guard accumulates `max_sstables_per_guard` number of sstables, the guard is compacted into the next level.

Tuning `max_sstables_per_guard` allows the user to trade-off more write IO (due to more compaction) for lower read and range query latencies. Interestingly, if this parameter is set to one, FLSM behaves like LSM and obtains similar read and write performance. Thus, FLSM can be viewed as a *generalization* of the LSM data structure.

3.6 Limitations

The FLSM data structure significantly reduces write amplification and has faster compaction (as compaction in FLSM requires lower read and write IO). By virtue of faster compaction, write throughput increases as well. However, the FLSM data structure is not without limitations.

Since `get()` and range query operations need to examine all sstables within a guard, the latency of these operations is increased in comparison to LSM. Chapter 4 describes how this limitation can be overcome; using a combination of well-known techniques can reduce or eliminate the overheads introduced by the FLSM data structure, resulting in a key-value store that achieves the trifecta of low write amplification, high write throughput, and high read throughput.

3.7 Asymptotic Analysis

This section provides an analysis of FLSM operations using a theoretical model.

Model. We use the standard Disk Access Model (DAM) [2] and assume that each read/write operation can access a block of size B in one unit cost. To simplify the model, we will assume a total of n data items are stored.

FLSM Analysis. Consider a FLSM where the guard probability is $1/B$ (so the number of guards in level $i+1$ is in expectation B times more than

the number of guards in level i). Since the expected fan-out of FLSM is B, with high probability, an FLSM with n data items will have $H = \log_B n$ levels. It is easy to see that each data item is written just once per level (it is appended once and never re-written to the same level), resulting in a write cost of $O(H) = O(\log_B n)$. Since in the DAM model, FLSM writes a block of B items at unit cost, the total amortized cost of any put operation is $O(H/B) = O((\log_B n)/B)$ over its entire compaction lifetime. However, FLSM compaction in the last level does re-write data. Since this last level re-write will occur with high probability $O(B)$ times then the final total amortized cost of any put operation is $O((B + \log_B n)/B)$.

The guards in FLSM induce a degree B Skip List. A detailed theoretical analysis of the B-Skip List data structure shows that with high probability each guard will have $O(B)$ children, each guard will have at most $O(B)$ sstables, and each sstable will have at most $O(B)$ data items [1, 24, 12]. Naively, searching for an item would require finding the right guard at each level (via binary search), and then searching inside all sstables inside the guard. Since the last level has the most guards (B^H), binary search cost would be dominated by the cost for the last level: $O(\log_2 B^H) = O(H \log_2 B) = O(\log_B n * \log_2 B) = O(\log_2 n)$. Since there are $O(H) = O(log_B n)$ levels to search, this yields a total cost of $O(\log_2 n \log_B n)$ in-memory operations for finding the right guards at each level.

However, in FLSM, the guards and bloom filters are all stored in memory. FLSM performs $O(\log_2 n \log_B n)$ in-memory operations during the binary

search for the right guard in each level. Then, for each of the $H = \log_B n$ guards found, FLSM does a bloom filter query on each of the $O(B)$ sstables associated with the guard, with each query costing $O(\log(1/\epsilon))$ in memory operations. In the DAM model all this in-memory work has no cost.

Finally, on average, the bloom filter will indicate only $1 + o(1)$ sstables to be read (with high probability). Reading these sstables will cost $1 + o(1)$ in the DAM model. Therefore, the total read cost of a get operation (assuming sufficient memory to store guards and bloom filters) is just $O(1)$ in the DAM model.

FLSM cannot leverage bloom filters for range queries. The binary search per level is still done in memory. For each level, the binary search outputs one guard and FLSM needs to read all the $O(B)$ associated sstables. So the total cost for a range query returning k elements is $O(B \log_B n + k/B)$.

Chapter 4

Building PebblesDB over FLSM

This chapter presents the design and implementation of PebblesDb, a high-performance key-value store built using fragmented log-structured merge trees. This chapter describes how PebblesDb offsets FLSM overheads for reads (section 4.1) and range queries (section 4.2), different PebblesDb operations (section 4.3), how PebblesDb recovers from crashes (section 4.3.1), its implementation (section 4.4), and its limitations (section 4.5).

4.1 Improving Read Performance

Overhead Cause. A `get()` operation in FLSM causes all the sstables of one guard in each level to be examined. In contrast, in log-structured merge trees, exactly one sstable per level needs to be examined. Thus, read operations incur extra overhead in FLSM-based key-value stores.

Sstable Bloom Filters. A Bloom filter is a space-efficient probabilistic data structure used to test whether an element is present in a given set in constant time [13]. A bloom filter can produce false positives, but not false negatives. PebblesDb attaches a bloom filter to each sstable to efficiently detect if a given key could be present in the sstable. The sstable bloom filters allow Peb-

blesDb to avoid reading unnecessary sstables off storage and greatly reduces the read overhead due to the FLSM data structure.

RocksDB also employs sstable-level bloom filters. Many key-value stores (including RocksDB and LevelDB) employ bloom filters for each block of the sstable. If sstable-level bloom filters are used, block-level filters are not required.

4.2 Improving Range Query Performance

Overhead Cause. Similar to `get()` operations, range queries (implemented using `seek()` and `next()` calls) also require examining all the sstables of a guard for FLSM. Since LSM stores examine only one sstable per level, FLSM stores have significant overhead for range queries.

Seek-Based Compaction. Similar to LevelDB, PebblesDb implements compaction triggered by a threshold number of consecutive `seek()` operations (default: 10). Multiple sstables inside a guard are merged and written to the guards in the next level. The goal is to decrease the average number of sstables within a guard. PebblesDb also aggressively compacts levels: if the size of level i is within a certain threshold ratio (default: 25%) of level $i+1$, level i is compacted into level $i+1$. Such aggressive compaction reduces the number of active levels that need to be searched for a `seek()`. Although such compaction increases write IO, PebblesDb still does significantly lower amount of IO overall (chapter 5).

Parallel Seeks. A unique optimization employed by PebblesDb is using multiple threads to search sstables in parallel for a `seek()`. Each thread reads one sstable off storage and performs a binary search on it. The results of the binary searches are then merged and the iterator is positioned correctly for the `seek()` operation. Due to this optimization, even if a guard contains multiple sstables, FLSM `seek()` latency incurs only a small overhead compared to LSM `seek()` latency.

Parallel seeks must not be carelessly used: if the sstables being examined are cached, the overhead of using multiple threads is *higher* than the benefit obtained from doing parallel seeks. Given that there is no way to know whether a given sstable has been cached or not (since the operating system may drop a cached sstable under memory pressure), PebblesDb employs a simple heuristic: parallel seeks are used only in the last level of the key-value store. The reason to choose this heuristic is that the last level contains the largest amount of data; furthermore, the data in the last level is not recent, and therefore not likely to be cached. This simple heuristic seems to work well in practice.

4.3 PebblesDB Operations

This section briefly describes how various operations are implemented in PebblesDb, and how they differ from doing the same operations on the FLSM data structure. The `put()` operation in PebblesDb is handled similar to puts in FLSM.

Get. PebblesDb handles `get()` operations by locating the appropriate guard in each level (via binary search) and searching the sstables within the guard. PebblesDb `get()` differs from FLSM `get()` in the use of sstable-level bloom filters to avoid reading unnecessary sstables off storage.

Range Query. PebblesDb handles range queries by locating the appropriate guard in each level and placing the iterator at the right position for each sstable in the guard by performing binary searches on the sstables. PebblesDb optimizes this by reading and searching sstables in parallel, and aggressively compacting the levels if a threshold number of consecutive `seek()` requests are received.

Deleting Keys. PebblesDb deletes a key by inserting the key into the store with a flag marking it as deleted. The sequence number of inserted key identifies it as the most recent version of the key, instructing PebblesDb to discard the previous versions of the key for read and range query operations. Note that bloom filters are created over sstables; since sstables are never updated in place, existing bloom filters do not need to be modified during key deletions. Keys marked for deletion are garbage collected during compaction.

4.3.1 Crash Recovery

By only appending data, and never over-writing any data in place, PebblesDb builds on the same foundation as LSM to provide strong crash-consistency guarantees. PebblesDb builds on the LevelDB codebase, and

LevelDB already provides a well-tested crash-recovery mechanism for both data (the sstables) and the metadata (the `MANIFEST` file). PebblesDb simply adds more metadata (guard information) to be persisted in the `MANIFEST` file. PebblesDb sstables use the same format as LevelDB sstables. Crash-recovery tests (testing recovered data after crashing at randomly picked points) confirm that PebblesDb recovers inserted data and associated guard-related metadata correctly after crashes.

4.4 Implementation

PebblesDb is implemented as a variant of the LevelDB family of key-value stores. PebblesDb was built by modifying HyperLevelDB [29], a variant of LevelDB that was engineered to have improved parallelism and better write throughput during compaction. We briefly examined the RocksDB code base, but found that the HyperLevelDB code base was smaller, better documented (as it derives from LevelDB), and easier to understand. Thus, HyperLevelDB was chosen as the base for PebblesDb.

We added/modified 9100 LOC in C++ to HyperLevelDB. Most of the changes involved introducing guards in HyperLevelDB and modifying compaction. Since guards are built *on top of* sstables, PebblesDb was able to take advantage of the mature, well-tested code that handled sstables. PebblesDb is API-compatible with HyperLevelDB since all changes are internal to the key-value store.

Selecting Guards. Similar to skip lists, PebblesDb picks guards randomly out of the inserted keys. When a key is inserted, a random number is selected to decide if the key is a guard. However, obtaining a random number for every key insertion is computationally expensive; instead, PebblesDb hashes every incoming key, and the last few bits of the hash determine if the key will be a guard (and at which level).

The computationally cheap `MurmurHash` [8] algorithm is used to hash each inserted key. A configurable parameter `top_level_bits` determines how many consecutive Least Significant Bits (LSBs) in the bit representation of the hashed key should be set for the key to be selected as a guard key in Level 1. Another parameter `bit_decrement` determines the number of bits by which the constraint (number of LSBs to be set) is relaxed going each level higher. For example, if `top_level_bits` is set to 17, and `bit_decrement` is set to 2, then a guard key in level 1 should have 17 consecutive LSBs set in its hash value, a guard key in level 2 should have 15 consecutive LSBs set in its hash value and so on. The `top_level_bits` and `bit_decrement` parameters need to be determined empirically; based on our experience, a value of two seems reasonable for `bit_decrement`, but the `top_level_bits` may need to be increased from our default of 27 if the users expect more than 100 million keys to be inserted into PebblesDb. Over-estimating the number of keys in the store is harmless (leads to many empty guards); under-estimating could lead to skewed guards.

Implementing Guards. Each guard stores metadata about the number of

sstables it has, the largest and smallest key present across the sstables, and the list of sstables. Each sstable is represented by a unique 64-bit integer. Guards are persisted to storage along with metadata about the sstables in the key-value store. Guards are recovered after a crash from the `MANIFEST` log and the asynchronous write-ahead logs. Recovery of guard data is woven into the key-value store recovery of keys and sstable information. The guard deletion is not implemented in PebblesDb yet since extra guards did not cause significant performance degradation for reads in our experiments and the cost of persisting empty guards is relatively insignificant, and the guard deletion can be added as part of the future work.

Multi-threaded Compaction. Similar to RocksDB, PebblesDb uses multiple threads for background compaction. Each thread picks one level and compacts it into the next level. Picking which level to compact is based on the amount of data in each level. When a level is compacted, only guards containing more than a threshold number of sstables are compacted. The guard-based parallel compaction is not implemented in PebblesDb yet; even without parallel compaction, compaction in PebblesDb is much faster than compaction in LSM-based stores such as RocksDB (section 5.2). Adding guard-based parallel compaction (as part of future work) will make the compaction in PebblesDb even more faster.

4.5 Limitations

This section describes three situations where a traditional LSM-based store may be a better choice over PebblesDb.

First, if the workload data will fit entirely in memory, PebblesDb has higher read and range query latency than LSM-based stores. In such a scenario, read or range query requests will not involve storage IO and the computational overhead of locating the correct guard and processing sstables inside a guard will contribute to higher latency. Given the increasing amount of data being generated and processed every day [50], most datasets will *not* fit in memory. For the rare cases where the data size is small, setting `max_sstables_per_guard` to one configures PebblesDb to behave similar to HyperLevelDB, reducing the latency overhead for reads and range queries.

Second, for workloads where data with sequential keys is being inserted into the key-value store, PebblesDb has higher write IO than LSM-based key value stores. If data is inserted sequentially, sstables don't overlap with each other. LSM-based stores handle this case efficiently by simply *moving* an sstable from one level to the next by modifying only the metadata (and without performing write IO); in the case of PebblesDb, the sstable may be partitioned when moving to the next level, leading to write IO. We believe that real-world workloads that insert data sequentially are rare since most workloads are multi-threaded; in such rare cases, we advocate the use of LSM-based stores such as RocksDB.

Third, if the workload involves an initial burst of writes followed by a large number of small range queries, PebblesDb may not be the best fit. For such range queries over a compacted key-value store, PebblesDb experiences a significant overhead (30%) compared to LSM-based stores. However, the overhead drops as the range queries get bigger and entirely disappears if the range queries are interspersed with insertions or updates (as in YCSB Workload E).

Chapter 5

Evaluation

This chapter evaluates the performance of PebblesDb by answering the following questions:

- What is the write amplification of PebblesDb? (section 5.2) What is the performance of various PebblesDb key-value store operations? (section 5.2) What are the strengths and weaknesses of PebblesDb?
- How does PebblesDb perform on workloads resembling access patterns in various applications? (section 5.3)
- How do NoSQL applications perform when they use PebblesDb as their storage engine? (section 5.4)
- How much memory and CPU does PebblesDb consume? (section 5.5)

5.1 Experimental Setup

Our experiments are run on a Dell Precision Tower 7810 with an Intel Xeon 2.8 GHz processor, 16 GB RAM, and running Ubuntu 16.04 LTS with the Linux 4.4 kernel. The ext4 file system is run on top of a software RAID0 array used over two high-performance Intel 750 SSDs (each 1.2 TB).

All workloads use datasets 3× larger than the main memory on test machine. All reported numbers are the mean of at least five runs. The standard deviation in all cases was less than 5% of the mean. PebblesDb performance is compared with widely-used key-value stores LevelDB, RocksDB and HyperLevelDB. To simplify results, compression is turned off in all stores. We have verified that compression does not change any of our performance results; it simply leads to a smaller dataset. HyperLevelDB does not employ bloom filters for sstables; to make a fair comparison (and to show our results do not derive just from sstable bloom filters), this optimization is added to HyperLevelDB: all numbers presented for HyperLevelDB are with bloom filters for sstables.

Key-Value Store Configurations. The key-value stores being evaluated have three configuration parameters that affect performance: `memtable-size`, `level0-slowdown`, `level0-stop`. Note that Level 0 can have sstables with overlapping ranges; new sstables are simply appended to Level 0 (otherwise adding an sstable to Level 0 would trigger compaction, affecting write throughput). However, letting Level 0 grow without bounds will reduce read and range query throughput. The `memtable-size` parameter controls how big the memtable can grow before being written to storage. The other two parameters are used to slow down or stop writes to Level 0.

HyperLevelDB and RocksDB have different default values for these parameters. HyperLevelDB uses 4 MB memtables with `level0-slowdown` of 8 and `level0-stop` of 12. RocksDB uses 64 MB memtables, `level0-slowdown`

of 20, and `level0-stop` of 24. When comparing PebblesDb with these systems, the default HyperLevelDB parameters are used. Certain experiments also report performance under RocksDB parameters.

5.2 Micro-benchmarks

This section evaluates PebblesDb performance using different single-threaded and multi-threaded micro-benchmarks and in various conditions. The single-threaded benchmarks help us understand the performance of different PebblesDb operations. The multi-threaded benchmark evaluates how PebblesDb performs in the more realistic setting of multiple readers and writers. PebblesDb is evaluated in different conditions such as when the dataset fits in memory, with small key-value pairs, with an aged file system and key-value store, and finally under extremely low memory conditions.

Write Amplification. We measure write amplification for workloads that insert or update keys in random order (key:16 bytes, value:128 bytes). Figure 5.1 (a) presents the results. PebblesDb write IO (in GB) is shown over the bars. PebblesDb consistently writes the least amount of IO, and the difference in write amplification between PebblesDb and other stores goes up as the number of keys increases. For 500M keys, PebblesDb lowers write amplification by 2.5× compared to RocksDB and HyperLevelDB and 1.6× compared to LevelDB.

Single-threaded Workloads. We use `db_bench` (a suite of micro-benchmarks

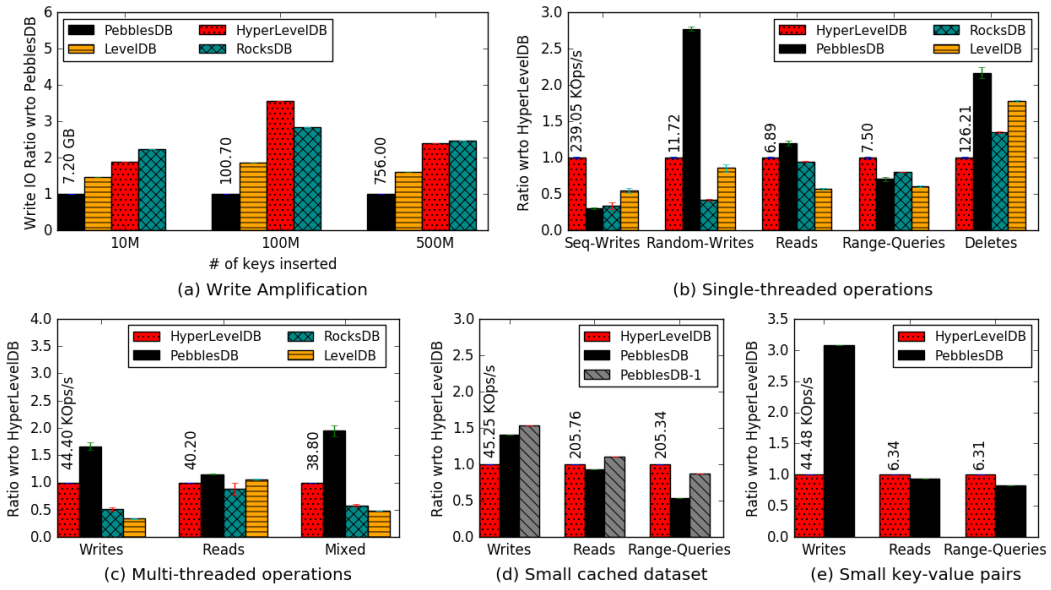


Figure 5.1: **Micro-benchmarks**. The figure compares the throughput of several key-value stores on various micro-benchmarks. Values are shown relative to HyperLevelDB, and the absolute value (in KOps/s or GB) of the baseline is shown above the bar. For (a), lower is better. In all other graphs, higher is better. PebblesDb excels in random writes, achieving 2.7× better throughput, while performing 2.5× lower IO.

that comes bundled with LevelDB) [33] to evaluate PebblesDb performance on various operations: 50M sequential writes, 50M random writes, 10M random reads and 10M random seeks. Reads and seeks were performed on the previously (randomly) inserted 50M keys. Each key was 16 bytes and the value was 1024 bytes. The results, presented in Figure 5.1 (b), show both the strengths and weaknesses of PebblesDb.

Random Writes and Reads. PebblesDb outperforms all other key-value stores in random writes due to the underlying FLSM data structure. PebblesDb throughput is 2.7× that of HyperLevelDB, the closest competitor.

	PebblesDB	HyperLevelDB
Average	17.23	13.33
Median	5.29	16.59
90th percentile	51.06	16.60
95th percentile	68.31	16.60

Table 5.1: **SSTable Size**. The table shows the distribution of sstable size (in MB) for PebblesDB and HyperLevelDB when 50 million key-value pairs totaling 33 GB were inserted.

PebblesDb compaction finishes 2.5× faster than HyperLevelDB compaction. Random reads perform better in PebblesDb due to the larger sstables of PebblesDb (as shown in Table 5.1). The index blocks of all PebblesDb sstables are cached, whereas there are cache misses for the index blocks of the many HyperLevelDB sstables. With larger caches for index blocks, PebblesDb read performance is similar to HyperLevelDB.

Sequential Writes. PebblesDb obtains 3× less throughput than HyperLevelDB on the sequential write workload; this is because sequential workloads result in disjoint sstables naturally (*e.g,.* first 100 keys go to the first sstable, next 100 keys go to the second sstable, *etc.*), LSM-based stores can just *move* the sstable from one level to another without doing any IO. On the other hand, PebblesDb always has to partition sstables based on guards (and therefore perform write IO) when moving sstables from one level to the next. As a result, PebblesDb performs poorly when keys are inserted sequentially. Many real-world workloads are multi-threaded, resulting in random writes; for example, in the YCSB workload suite which reflects real-world access patterns, none of the workloads insert keys sequentially [16].

Range Queries. A range query is comprised of an `seek()` operation followed by a number of `next()` operations. Range-query performance depends mainly on two factors: the number of levels in the key-value store on storage, and the number of `next()` operations. Figure 5.1 (b) shows key-value store performance for range queries comprising of only `seek()` operations, performed after allowing the key-value store time to perform compaction. As such, it represents a worst case for PebblesDb: the expensive `seek()` operation is not amortized by successive `next()` operations, and other key-value stores compact more aggressively than PebblesDb, since they do not seek to minimize write IO. In this worst-case scenario, PebblesDb has a 30% overhead compared to HyperLevelDB, due to the fact that a `seek()` in PebblesDb requires reading multiple sstables from storage in each level. We note that in real-world workloads such as YCSB, there are many `next()` operations following a `seek()` operation.

Next, we measure range query performance in a slightly different setting. We insert 50M key-value pairs (key: 16 bytes, value: 1 KB), and immediately perform 10M range queries (each range query involves 50 `next()` operations). In this more realistic scenario, we find that PebblesDb overhead (as compared to HyperLevelDB) reduces to 15% from the previous 30%. If we increase range query size to 1000, the overhead reduces to 11%.

Unfortunately, even with many `next()` operations, PebblesDb range-query performance will be lower than that of LSM key-value stores. This is because PebblesDb pays both an IO cost (reads more sstables) and a CPU

	PebblesDB	HyperLevelDB	LevelDB	RocksDB
Insert 50M values	56.18	40.00	22.42	14.12
Update Round 1	47.85	24.55	12.29	7.60
Update Round 2	42.55	19.76	11.99	7.36

Table 5.2: **Update Throughput**. The table shows the throughput in KOps/s for inserting and updating 50M key-value pairs in different key-value stores.

cost (searches through more sstables in memory, merges more iterators) for range queries. While the overhead will drop when the number of `next()` operations increase (as described above), it is difficult to eliminate both IO cost and CPU cost.

To summarize range-query performance, PebblesDb has significant overhead (30%) for range queries when the key-value store has been fully compacted. This overhead derives both from the fact that PebblesDb has to examine more sstables for a `seek()` operation, and that PebblesDb does not compact as aggressively as other key-value stores as it seeks to minimize write IO. The overhead is reduced for large range queries, and when range queries are interspersed with writes (such as in YCSB workload E).

Deletes and Updates. Deletes and Updates are handled similar to writes in LSM-based key-value stores. Updates do not check for the previous value of the key, so updates and new writes are handled identically. Deletes are simply writes with a zero-sized value and a special flag. We ran an experiment where we inserted 200M key-value pairs (key: 16 bytes, value: 128 bytes) into the database and deleted all inserted keys. We measure the deletion throughput.

The results are presented in Figure 5.1 (b) and follow a pattern similar to writes: PebblesDb outperforms the other key-value stores due to its faster compaction.

We ran another experiment to measure update throughput. We inserted 50M keys (value: 1024 bytes) into the store, and then updated all keys twice. The results are presented in Table 5.2. We find that as the database becomes larger, insertion throughput drops since insertions are stalled by compactions and compactions involve more data in larger stores. While the other key-value stores drop to 50% of the initial write throughput, PebblesDB drops to only 75% of original throughput; we attribute this difference to the compaction used by the different key-value stores. The update throughput of PebblesDB is 2.15× that of HyperLevelDB, the closest competitor.

Multi-threaded Reads and Writes. We use four threads to perform 10M read and 10M write operations (each) on the evaluated key-value stores. The reads are performed on the store after the write workload finishes. We use the default RocksDB configuration (64 MB memtable, large Level 0). Figure 5.1 (c) presents the results. PebblesDb performs the best on both workloads, obtaining 3.3× the write throughput of RocksDB (1.7× over baseline).

Concurrent Reads and Writes. In this experiment, two threads perform 10M reads each, while two other threads perform 10M writes each. Figure 5.1 (c) reports the combined throughput of reads and writes (*mixed*). PebblesDb outperforms the other stores. The lower write amplification leads to higher

write throughput. Since compaction in PebblesDb is faster than the other stores, PebblesDb reaches a compacted state earlier with larger (and fewer) sstables, resulting in lower read latency and higher read throughput. Note that PebblesDb outperforms HyperLevelDB even when HyperLevelDB uses sstable-level bloom filters, thus demonstrating the gains are due to the underlying FLSM data structure.

Small Workloads on Cached Datasets. We run an experiment to determine the performance of PebblesDb on data sets that are likely to be fully cached. We insert 1M random key-value pairs (key:16 bytes, value: 1KB) into HyperLevelDB and PebblesDb. The total dataset size is 1 GB, so it is comfortably cached by the test machine (RAM: 16 GB). We do 1M random reads and seeks. Figure 5.1 (d) presents the results. Even for small datasets, PebblesDb gets better write throughput than HyperLevelDB due to the FLSM data structure. Due to extra CPU overhead of guards, there is a small 7% overhead on reads and 47% overhead on seeks. When PebblesDb is configured to run with `max_sstables_per_guard` (section 3.5) set to one so that it behaves more like an LSM store (*PebblesDB-1*), PebblesDb achieves 11% higher read throughput and the seek overhead drops to 13%.

Performance for Small Sized Key-Value Pairs. We insert 300M key-value pairs into the database (key: 16 bytes, value: 128 bytes). As shown in Figure 5.1 (e), PebblesDb obtains higher write throughput and equivalent read and seek throughputs (similar to results with large keys).

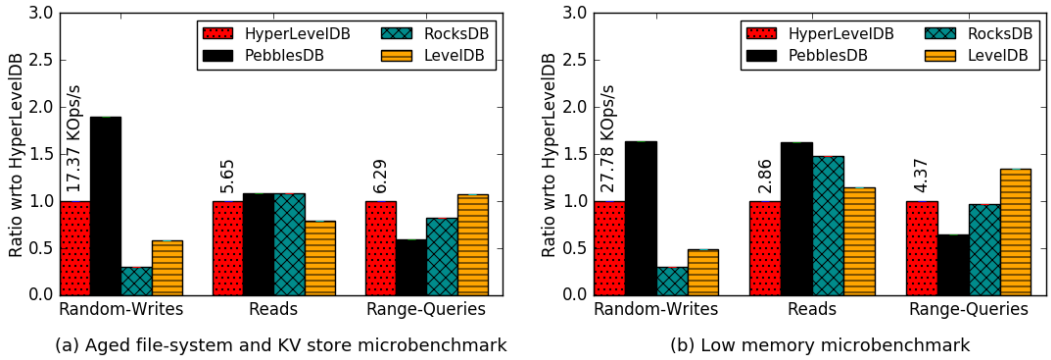


Figure 5.2: **Effect of environmental parameters**. The figure compares the throughput of several key-value stores on varying the environmental parameters like the age of file system (or key-value store), and the amount of memory available. Values are shown relative to HyperLevelDB, and the absolute value (in KOps/s) of the baseline is shown above the bar. The higher is better.

Impact of File-System and Key-Value Store Aging. Recent work has shown that file-system aging has a significant impact on performance [15]. To assess the impact of file-system and key-value store aging on PebblesDb, we run the following experiment. *File-system Aging*: We create a new file system on a 1.1 TB SSD, then use sequential key-value pair insertion to fill up the file system. We then delete all data in the file system, and fill the file system using the same process again until 130 GB of free space (11% of the file-system size) is left. *Key-Value Store Aging*: We then age the key-value store under evaluation by using four threads to each insert 50M key-value pairs, delete 20M key-value pairs, and update 20M key-value pairs in random order. Once both file-system and key-value store aging is done, we then run micro-benchmarks for writes, reads, and seeks (all in random order). The results are presented in Figure 5.2 (a). We find that the absolute performance numbers drop: 18% for reads and 16% for range queries (mainly because there is more data in the key-value

store from the aging). As with a fresh file system, PebblesDb outperforms the other key-value stores on writes (although the throughput speedup reduces to 2× from 2.7×). Similarly, PebblesDb outperforms HyperLevelDB by 8% (down from 20% on a fresh file system) on reads, and incurs a 40% penalty on range queries (as compared to 30% on a fresh file system) compared to HyperLevelDB.

Performance Under Low Memory. We evaluate the performance of PebblesDb when the total available memory is a small percentage of the dataset size. We insert 100M key-value pairs (key:16 bytes, value: 1K) for a total dataset size of 65 GB. We restrict the RAM on our machine using the `mem` kernel boot parameter to 4 GB. Thus, the total available DRAM is only 6% of the total dataset size (in our previous experiments, it was 30%). We evaluate the performance of PebblesDb under these conditions using micro-benchmarks. The results are presented in Figure 5.2 (b). All key-values stores evaluated use a 64 MB memtable and a large Level 0 (with `level0-slowdown` as 20 and `level0-stop` as 24). We find that PebblesDb still outperforms the other key-value stores at random writes, although the margin (with respect to HyperLevelDB) reduces to 64%. PebblesDb outperforms HyperLevelDB on random reads by 63%. On the range query micro-benchmark, PebblesDb experiences a 40% penalty compared to HyperLevelDB. Thus, PebblesDb still achieves good performance in reads and writes when memory is scarce, although range queries experience more performance degradation.

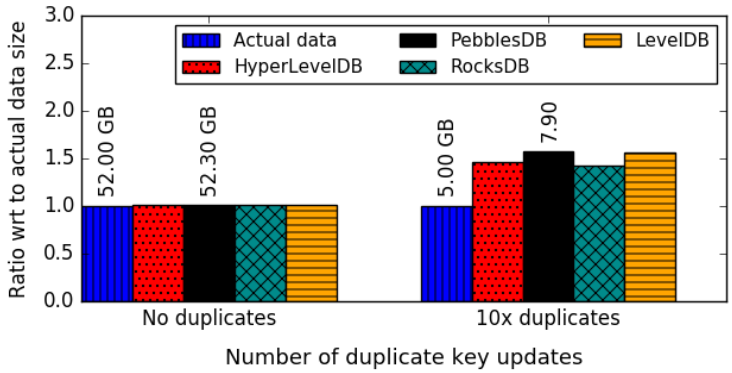


Figure 5.3: **Space amplification**. The figure compares the space amplification of the different key-value stores. Values are shown relative to the actual data size and the absolute value (in GB) of the baseline is shown above the bar. PebblesDb doesn't incur high space amplification compared to the other key-value stores even on inserting 10x times duplicate keys.

Space Amplification. The storage space used by PebblesDb is not significantly higher compared to LSM-based stores. LSM-based stores only reclaim space if the key has been updated or deleted. For a workload with only insertions of unique keys, the space used by RocksDB and PebblesDB will be identical. For workloads with updates and deletions, PebblesDB will have a slight overhead due to delay in merging. Figure 5.3 shows the results of experiments on space amplification. We inserted 50M unique key-value pairs. The storage-space consumption of RocksDB, LevelDB, and PebblesDb were within 2% of each other (52 GB). We performed another experiment where we inserted 5M unique keys, and updated each key 10 times (total 50M writes). Since the keys aren't compacted yet, PebblesDb consumes 7.9 GB while RocksDB consumes 7.1 GB. LevelDB consumed 7.8 GB of storage space.

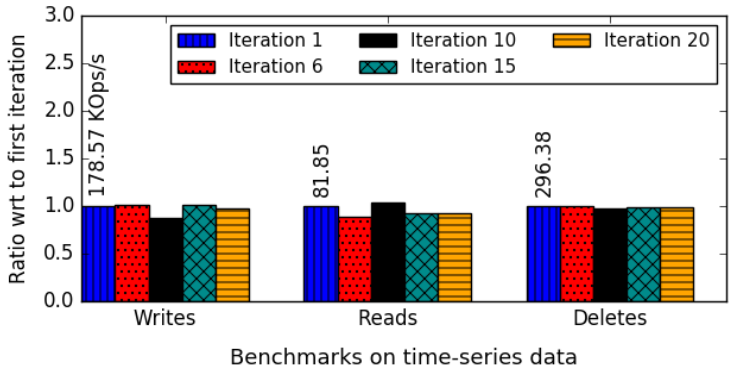


Figure 5.4: **Time-series data**. The figure compares the throughput of the different key-value stores on time-series data (to measure impact of empty guards). Values are shown relative to the to the first iteration, and the absolute value (in KOps/s) of the baseline is shown above the bar. The performance of PebblesDb is unaffected by time-series pattern of the data.

Impact of Empty Guards. We run an experiment to measure the performance impact of empty guards. We insert 20M key-value pairs (with keys from 0 to 20M, value size: 512B, dataset size: 10 GB), perform 10M read operations on the data, and delete all keys. We then repeat this, but with keys from 20M to 40M. We do twenty iterations of this experiment and the results are presented in Figure 5.4. Since we are always reading the currently inserted keys, empty guards due to old deleted keys will accumulate (there are 9000 empty guards at the beginning of the last iteration). Throughout the experiment, read throughput varied between 70 and 90 KOps/s. Read throughput did not reduce with more empty guards and so write and delete throughputs.

Impact of Different Optimizations. We briefly describe how the different optimizations described in the thesis affect PebblesDb performance. If Peb-

Workload	*Description*	*Represents*
Load A	100% writes	Insert data for workloads A–D and F
A	50% reads, 50% writes	Session recording recent actions
B	95% reads, 5% writes	Browsing and tagging photo album
C	100% reads	Caches
D	95% reads (latest values), 5% writes	News feed or status feed
Load E	100% writes	Insert data for Workload E
E	95% Range queries, 5% writes	Threaded conversation
F	50% reads, 50% Read-modify-writes	Database workload

Table 5.3: **YCSB Workloads**. The table describes the six workloads in the YCSB suite. Workloads A–D and F are preceded by Load A, while E is preceded by Load E.

bLESDb doesn't use any optimizations for range queries, range query throughput drops by 66% (48 GB dataset). The overhead drops to 48% if parallel seeks are used, and only 7% if only seek-based compaction is used. Using sstable-level bloom filters improves read performance by 63% (53 GB dataset).

5.3 Yahoo Cloud Serving Benchmark

The industry standard in evaluating key-value stores is the Yahoo Cloud Serving Benchmark [16]. The suite has six workloads (described in Table 5.3), each representing a different real-world scenario. We modify `db_bench` [33]

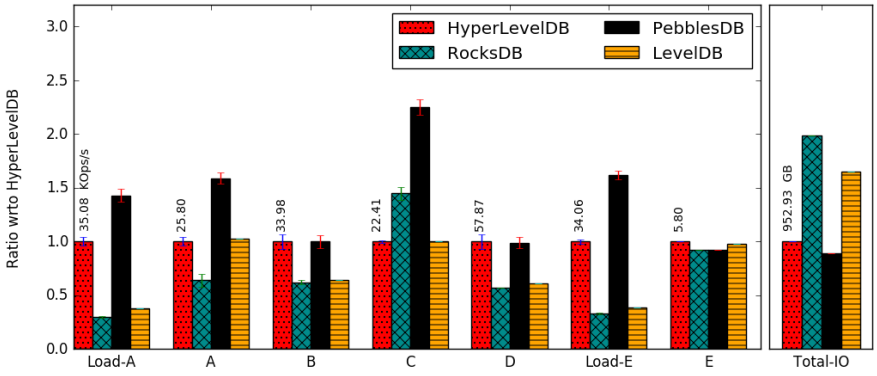


Figure 5.5: **YCSB Performance**. The figure shows the throughput (bigger is better except for Total-IO bars) of different key-value stores on the YCSB Benchmark suite run with four threads. PebblesDb gets higher throughput than RocksDB on almost all workloads, while performing 2× lower IO than RocksDB.

to run the YCSB benchmark with 4 threads (one per core) and using default RocksDB parameters (64MB memtable and large Level 0). We run RocksDB with 4 background compaction threads to further boost its performance. Load-A and Load-E do 50M operations each, all other workloads do 10M operations each. Figure 5.5 presents the results: PebblesDb outperforms both RocksDB and HyperLevelDB on write workloads, while obtaining nearly equal performance on all other workloads. Overall, PebblesDb writes 50% less IO than RocksDB.

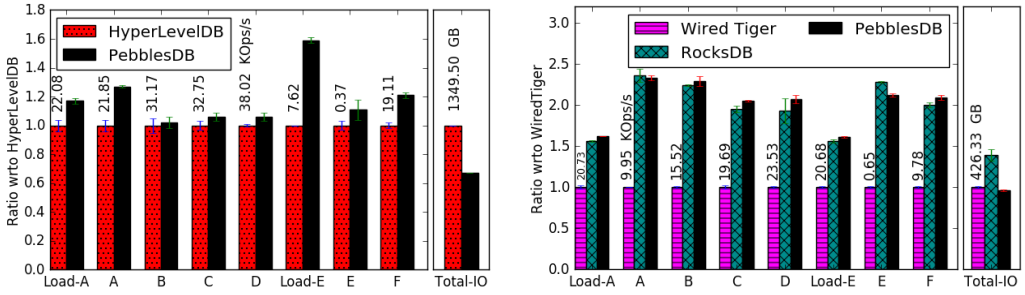
On write-dominated workloads like Load A and Load E, PebblesDb achieves 1.5–2× better throughput due to the faster writes offered by the underlying FLSM data structure.

For the read-only Workload C, PebblesDb read performance is better than other key-value stores due to the larger sstables of PebblesDb. The

key-value stores cache a limited number of sstable index blocks (default: 1000): since PebblesDb has fewer, larger files, most of its sstable-index-blocks are cached. The cache misses for the other key-value stores result in reduced read performance. When we increase the number of index blocks cached, PebblesDb read performance becomes similar to the other key-value stores. Note that the larger sstables of PebblesDb result from compaction: in workloads such as B and D, the constant stream of writes adds new sstables that are not compacted; as a result, PebblesDb throughput is similar to the other key-value stores.

For the range-query-dominated Workload E, PebblesDb surprisingly has performance close (6% overhead) to the other key-value stores. When we analyzed this, we found that the small amount of writes in the workload (Workload E has 5% writes) prevent any key-value store from full compacting; as a result, every key-value store has to examine multiple levels, which reduces the performance impact of the extra sstables examined by PebblesDb. When the YCSB workload is modified to contain only range queries, PebblesDb throughput is 18% lower than HyperLevelDB as expected. Each range query in this workload does N `next()` operations (N picked randomly from 1 to 100), and the `next()` operations also contribute in reducing range-query overhead.

In Workload F, all writes are read-modify-writes: the workload does a `get()` before every `put()` operation. As a result, the full write throughput of PebblesDb is not utilized, resulting in performance similar to that of other key-value stores. We see similar read-modify-write behavior in applications



(a) HyperDex Throughput (b) MongoDB Throughput

Figure 5.6: **Application Throughput**. The figure shows the YCSB throughput (bigger is better except last bar) of the HyperDex document store and MongoDB NoSQL store when using different key-value stores as the storage engine. The throughput is shown relative to the default storage option (HyperLevelDB for HyperDex, WiredTiger for MongoDB). The raw throughput in KOps/s or total IO in GB of the default option is shown above the bars.

such as HyperDex and MongoDB.

5.4 NoSQL Applications

We evaluate the performance of two real-world applications, the HyperDex and MongoDB NoSQL stores, when they use PebblesDb as the underlying storage engine. We use the Java clients provided by HyperDex and MongoDB for running the YCSB benchmark, with both the server and client running on the same machine (no network involved).

HyperDex. HyperDex is a high-performance NoSQL store that uses HyperLevelDB as it storage engine by default [18]. We evaluate the performance impact of using PebblesDb as the storage engine by running the YCSB benchmark with 4 threads. Load-A inserts 20M values, Load-E inserts 30M

values, A–D and F perform 10M operations each, and E performs 250K operations (lower number of ops as HyperDex range-query latency is very high). We use the same setup used by HyperDex developers to benchmark their system using YCSB [28]. Both HyperLevelDB and PebblesDb are configured with the default HyperDex parameters (16 MB memtable size).

Figure 5.6 (a) presents the results. In every workload, using PebblesDb improves HyperDex throughput, with the highest gain of 59% coming when inserting 30M key-value pairs in the Load-E workload. HyperDex adds significant latency to operations done by YCSB. For example, the average latency to insert a key in HyperDex is 151 μs, of which PebblesDb accounts for only 22.3 μs (14.7%). Furthermore, HyperDex checks whether a key already exists before inserting, turning every `put()` operation in the Load workloads into a `get()` and a `put()`. This behavior of HyperDex reduces the performance gain from PebblesDb, because PebblesDb can handle much higher rate of insertions. Despite this, PebblesDb increases HyperDex throughput while simultaneously reducing write IO.

We measure the throughput loss that PebblesDb suffers as a storage backend to HyperDex because of the read-then-write behavior of HyperDex. We insert 5M key-value pairs to HyperDex; we measure the write throughput of both PebblesDb and HyperLevelDB on HyperDex for two scenarios - with the default read-then-write behavior, and by disabling reads before writes. For a value size of 8 KB, write throughput of PebblesDb is 7% better than HyperLevelDB while the same is 38% better on disabling the reads

before writes. For value size of 16 KB, write throughput of PebblesDb is 24% better than HyperLevelDB while the same is 110% better on disabling the reads before writes. This experiment shows that the throughput gain of PebblesDb compared to HyperLevelDB is limited by the read-then-write behavior of HyperDex. In general, PebblesDb is able to achieve higher write throughput compared to the other key-value stores if not limited by the latency introduced by the application itself.

When we increase the value size from the YCSB default of 1 KB to 16 KB, the speedup HyperDex achieves from using PebblesDb drastically increases: the geometric mean of the speedup is **105%** (not shown). As the value size increases, more IO is required for all operations, making the extra CPU overhead of PebblesDb negligent, and highlighting the benefits of the FLSM data structure.

MongoDB. We configure MongoDB [40], a widely-used NoSQL store, to use PebblesDb as the storage engine. MongoDB can natively run with either the Wired Tiger key-value store (default) or RocksDB. We evaluate all three options using the YCSB Benchmark suite. All three stores are configured to use 8 MB cache and a 16 MB memtable. Since Wired Tiger is not a LSM-based store (it uses checkpoints + journaling), it does not use memtables; instead, it collects entries in a log in memory. We configure the max size of this log to be 16 MB. Figure 5.6 (b) presents the results. We find that both RocksDB and PebblesDb significantly outperform Wired Tiger on all workloads, demonstrating why LSM-based stores are so popular. While RocksDB performs 40%

more IO than Wired Tiger, PebblesDb writes 4% lesser IO than Wired Tiger.

We investigated why PebblesDb write throughput is not 2× higher than RocksDB as in the YCSB benchmark. As in HyperDex, MongoDB itself adds a lot of latency to each write (PebblesDb write constitutes only 28% of latency of MongoDB write) and provides requests to PebblesDb at a much lower rate than PebblesDb can handle. The slower request rate allows RocksDB's compaction to keep up with the inserted data; thus, PebblesDb's faster compaction is not utilized, and the two key-value stores have similar write throughput. Note that PebblesDb still writes 40% lesser IO then RocksDB, providing lower write amplification.

Due to lack of time before the conference submission deadline, this particular evaluation section (evaluating PebblesDb on MongoDB) was done by my co-author, Rohan Kadekodi, Masters student at The University of Texas at Austin, and I helped him run the experiments.

Summary. PebblesDb does not increase performance on HyperDex and MongoDB as significantly as in the YCSB macro-benchmark. This is both due to PebblesDb latency being a small part of overall application latency, and due to application behavior such as doing a read before every write. If the application is optimized for PebblesDb, we believe the performance gains would be more significant. Despite this, PebblesDb reduces write amplification, providing either equal (MongoDB) or better performance (HyperDex).

Workload	*HyperLevelDB*	*RocksDB*	*PebblesDB*
Writes (100M)	159	896	434
Reads (10M)	154	36	500
Seeks (10M)	111	34	430

Table 5.4: **Memory Consumption**. The table shows the memory consumed (in MB) by key-value stores for different workloads.

5.5 Memory and CPU Consumption

Memory Consumption. We measure memory used during the insertion of 100M keys (key size: 16 bytes, value size: 1024 bytes, total: 106 GB) followed by 10M reads and range queries. The results are shown in Table 5.4. PebblesDb consumes about 300 MB more than HyperLevelDB. PebblesDb uses 150 MB for storing sstable bloom filters, and 150 MB for temporary storage for constructing the bloom filters.

CPU Cost. We measured the median CPU usage during the insertion of 30M keys, followed by reads of 10M keys. The median CPU usage of PebblesDb is 170.95%, while the median for the other key-value stores ranged from 98.3–110%. The increased CPU usage is due to the PebblesDb compaction thread doing more aggressive compaction.

Bloom Filter Construction Cost. Bloom filters are calculated over all the keys present in an sstable. The overhead of calculating the bloom filter is incurred only the first time the sstable is accessed. The time taken to calculate depends on the size of sstable. We observed the rate of calculation to be 1.2 s per GB of sstable; for 3200 sstables totaling 52 GB, it took around 62 seconds.

Chapter 6

Related Work

The work in this thesis builds on extensive prior work in building and optimizing key-value stores. The key contribution relative to prior work is the FLSM data structure and demonstrating that a high performance key-value store that drastically reduces write amplification can be built on top of FLSM. This chapter briefly describes prior work and places the work in this thesis in context.

Reducing Write Amplification. Various data structures have been proposed for implementing key-value stores. Fractal Index trees [11] (see TokuDB [36]) were suggested to reduce the high IO cost associated with traditional B-Trees. While FLSM and Fractal index trees share the same goal of reducing write IO costs, Fractal index trees do not achieve high write throughput by taking advantage of large sequential writes, and do not employ in-memory indexes such as bloom filters to improve performance like PebblesDb.

NVMKV [38] uses a hashing-based design to reduce write amplification and deliver close to raw-flash performance. NVMKV is tightly coupled to the SSD's Flash Translation Layer (FTL) and cannot function without using FTL features such as atomic multi-block write. Similarly, researchers have proposed

building key-value stores based on vector interfaces (that are not currently available) [56]. In contrast, PebblesDb is device-agnostic and reduces write amplification on both commodity hard drives and SSDs. We should note that we have not tested PebblesDb on hard-drives yet; we believe the write behavior will be similar, although range query performance may be affected.

The HB+-trie data structure is used in ForestDB [4] to efficiently index long keys and reduce space overhead of internal nodes. FLSM and HB+trie target different goals resulting in different design decisions; FLSM is designed to reduce write amplification, not space amplification.

The LSM-trie [58] data structure uses *tries* to organize keys, thereby reducing write amplification; however, it does not support range queries. Similarly, RocksDB's universal compaction reduces write amplification by sacrificing read and range query performance [22]. PebblesDb employs additional techniques over FLSM to balance reducing write amplification with reasonable range query performance.

TRIAD [10] uses a combination of different techniques such as separating hot and cold keys, using commit logs as sstables, and delaying compaction to reduce write IO and improve performance. The TRIAD techniques are orthogonal to our work and can be incorporated into PebblesDb.

Improving Key-Value store Performance. Both academia and industry have worked on improving the performance of key-value stores based on log-structured merge trees. PebblesDb borrows optimizations such as

sstable bloom filters and multi-threaded compaction from RocksDB. HyperLevelDB [29] introduces fine-grained locking and a new compaction algorithm that increases write throughput. bLSM [52] introduces a new merge scheduler to minimize write latency and maintain write throughput, and uses bloom filters to improve performance. VT-Tree [53] avoids unnecessary data copying for data that is already sorted using an extra level of indirection. WiscKey [37] improves performance by not storing the values in the LSM structure. LOCS [57] improves LSM compaction using the internal parallelism of open-channel SSDs. cLSM [23] introduces a new algorithm for increasing concurrency in LSM-based stores. We have a different focus from these work: rather than making LSM-based stores better, we introduce a better data structure, FLSM, and demonstrate that it can be used to build high performance key-value stores. Many of the techniques in prior work can be readily adapted for FLSM and PebblesDb.

Chapter 7

Future Work

Although PebblesDb performs better in many workloads compared to the other key-values stores (as discussed in chapter 5), it is not without limitations. This chapter outlines some of the shortcomings of PebblesDb and describes few possible methods to address the shortcomings, as part of the future work.

Optimizing Range Queries. One of the main challenges that PebblesDb faces is having to examine multiple sstables per level during a `get()` or a `range_query` operation leading to higher latency. PebblesDb employs some techniques like sstable level bloom filters and parallel seeks to optimize the `get()` and `range_query` operations but the overhead is still not eliminated completely. Recent work on Succinct Range Filter (SuRF) [60] introduces *Fast Succinct Trie* (FST) which can be used to optimize the range queries (similar to bloom filters for get queries).

Increasing the parallelism of compaction. The guards in FLSM are arranged in form of a skip-list: a guard in level i serves as a guard in all levels greater than i as well. Because of this property, the background compaction can be trivially parallelized at the granularity of a *guard* instead of *level*. Since

the write throughput also directly depends on the rate of compaction happening in the background, introducing a guard-based compaction will speed up the overall compaction rate and hence increase the write throughput even further.

Optimizing Memory utilization. PebblesDb has a higher overhead in terms of the amount of memory used since it stores all the sstable level bloom filters in memory. This also affects the performance of reads since it leads to more cache misses for the read operations (higher the memory used by sstable level bloom filters, lower the memory available to cache the user data). As part of the future work, the memory utilization can be optimized by storing the sstable level bloom filters on the storage along with the sstables' index blocks instead of storing in memory and the filters can be fetched on demand from the storage. Although this results in extra overhead of reading the bloom filters from storage, a cache of sstable level bloom filters can be maintained in memory to reduce the number of storage reads (to fetch bloom filters) for frequently and recently used sstables.

Making Guards dynamic and adaptive. The guards in FLSM are statically determined probabilistically during the insertion of keys and the guards are not deleted currently in PebblesDb since having empty guards doesn't affect the performance. Empty guards can be deleted periodically to have cleaner and lighter meta-data. Static selection of guards can also lead to skew in the amount of data that is distributed between the guards. As part of future work, the selection of guards can be made dynamic and the data can

be re-balanced between the guards during background compaction (by adding new guards or deleting existing guards, thereby making the guards adaptive to the distribution of data) which will make PebblesDb more robust against any kind of data distribution.

Bibliography

[1] Ittai Abraham, James Aspnes, and Jian Yuan. Skip B-trees. In *Proceedings of the 9th International Conference on Principles of Distributed Systems (OPODIS 2005)*, pages 366–380, December 2005.

[2] Alok Aggarwal, Jeffrey Vitter, et al. The input/output complexity of sorting and related problems. *Communications of the ACM*, 31(9):1116–1127, 1988.

[3] Nitin Agrawal, Vijayan Prabhakaran, Ted Wobber, John D. Davis, Mark S. Manasse, and Rina Panigrahy. Design tradeoffs for ssd performance. In *Proceedings of the 2008 USENIX Annual Technical Conference*, pages 57–70, 2008.

[4] Jung-Sang Ahn, Chiyoung Seo, Ravi Mayuram, Rahim Yaseen, Jin-Soo Kim, and Seungryoul Maeng. Forestdb: A fast key-value storage system for variable-length string keys. *IEEE Transactions on Computers*, 65(3):902–915, 2016.

[5] Reed Allman. Rock solid queues @ iron.io. `https://www.youtube.com/watch?v=HTjt6oj-RL4`, 2014.

[6] David G. Andersen, Jason Franklin, Michael Kaminsky, Amar Phanishayee, Lawrence Tan, and Vijay Vasudevan. Fawn: A fast array of

wimpy nodes. In *Poceedings of the ACM SIGOPS 22nd Symposium On Operating Systems Principles (SOSP 09)*, pages 1–14. ACM, 2009.

[7] Apache. Search results apache flink: Scalable stream and batch data processing. `https://flink.apache.org`, 2017.

[8] Austin Appleby. SMHasher test suite for MurmurHash family of hash functions. `https://github.com/aappleby/smhasher`, 2016.

[9] Anirudh Badam, KyoungSoo Park, Vivek S. Pai, and Larry L Peterson. Hashcache: Cache storage for the next billion. In *Proceedings of the 6th USENIX Symposium on Network Systems Design and Implementation (NSDI 09)*, pages 123–136, 2009.

[10] Oana Balmau, Diego Didona, Rachid Guerraoui, Willy Zwaenepoel, Huapeng Yuan, Aashray Arora, Karan Gupta, and Pavan Konka. TRIAD: Creating synergies between memory, disk and log in log structured key-value stores. In *Proceedings of the 2017 USENIX Annual Technical Conference (USENIX ATC 17)*, pages 363–375, Santa Clara, CA, 2017.

[11] Michael A. Bender, Martin Farach-Colton, Jeremy T. Fineman, Yonatan R. Fogel, Bradley C. Kuszmaul, and Jelani Nelson. Cache-oblivious streaming b-trees. In *Proceedings of the 19th Annual ACM Symposium on Parallel Algorithms and Architectures*, pages 81–92. ACM, 2007.

[12] Michael A. Bender, Martín Farach-Colton, Rob Johnson, Simon Mauras, Tyler Mayer, Cynthia A. Phillips, and Helen Xu. Write-optimized skip

lists. In *Proceedings of the 36th ACM Symposium on Principles of Database Systems*, PODS '17, pages 69–78, New York, NY, USA, 2017. ACM.

[13] Burton H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970.

[14] Douglas Comer. Ubiquitous b-tree. *ACM Computing Surveys (CSUR)*, 11(2):121–137, 1979.

[15] Alexander Conway, Ainesh Bakshi, Yizheng Jiao, William Jannen, Yang Zhan, Jun Yuan, Michael A. Bender, Rob Johnson, Bradley C. Kuszmaul, Donald E. Porter, Jun Yuan, and Martin Farach-Colton. File systems fated for senescence? nonsense, says science! In *Proceedings of the 15th USENIX Conference on File and Storage Technologies (FAST 17)*, pages 45–58, 2017.

[16] Brian F. Cooper, Adam Silberstein, Erwin Tam, Raghu Ramakrishnan, and Russell Sears. Benchmarking Cloud Serving Systems with YCSB. In *Proceedings of the 1st ACM Symposium on Cloud Computing (SOCC 10)*, pages 143–154. ACM, 2010.

[17] Biplob Debnath, Sudipta Sengupta, and Jin Li. Skimpystash: Ram space skimpy key-value store on flash-based storage. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, pages 25–36. ACM, 2011.

[18] Robert Escriva, Bernard Wong, and Emin Gün Sirer. Hyperdex: a distributed, searchable key-value store. In *Proceedings of the ACM SIGCOMM 2012 Conference*, pages 25–36, 2012.

[19] Facebook. Fifo compaction style. `https://github.com/facebook/rocksdb/wiki/FIFO-compaction-style`, 2017.

[20] Facebook. RocksDB — A persistent key-value store. `http://rocksdb.org`, 2017.

[21] Facebook. Rocksdb users. `https://github.com/facebook/rocksdb/blob/master/USERS.md`, 2017.

[22] Facebook. Universal compaction. `https://github.com/facebook/rocksdb/wiki/Universal-Compaction`, 2017.

[23] Guy Golan-Gueta, Edward Bortnikov, Eshcar Hillel, and Idit Keidar. Scaling Concurrent Log-structured Data Stores. In *Proceedings of the Tenth European Conference on Computer Systems (Eurosys 15)*, page 32. ACM, 2015.

[24] Daniel Golovin. The B-skip-list: A simpler uniquely represented alternative to B-trees. *CoRR*, abs/1005.0662, 2010.

[25] Google. Leveldb. `https://github.com/google/leveldb`, 2017.

[26] Laura M. Grupp, Adrian M. Caulfield, Joel Coburn, Steven Swanson, Eitan Yaakobi, Paul H. Siegel, and Jack K. Wolf. Characterizing flash

memory: Anomalies, observations, and applications. In *Proceedings of 42nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-42)*, pages 24–33. IEEE, 2009.

[27] James Hamilton. The cost of latency. `http://perspectives.mvdirona.com/2009/10/the-cost-of-latency/`, 2009.

[28] HyperDex. HyperDex Benchmark Setup. `http://hyperdex.org/performance/setup/`, 2016.

[29] HyperDex. Hyperleveldb performance benchmarks. `http://hyperdex.org/performance/leveldb/`, 2017.

[30] Cockroach Labs. Cockroachdb. `https://github.com/cockroachdb/cockroach`, 2017.

[31] Dgraph labs. Dgraph: Graph database for production environment. `https://dgraph.io`, 2017.

[32] FAL Labs. Kyoto Cabinet: a straightforward implementation of DBM. `http://fallabs.com/kyotocabinet/`, 2011.

[33] LevelDB. LevelDB db_bench benchmark. `https://github.com/google/leveldb/blob/master/db/db_bench.cc`, 2016.

[34] Hyeontaek Lim, Bin Fan, David G. Andersen, and Michael Kaminsky. Silt: A memory-efficient, high-performance key-value store. In *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles (SOSP 11)*, pages 1–13. ACM, 2011.

[35] LinkedIn. Followfeed: Linkedin's feed made faster and smarter. http://bit.ly/2onMQwN, 2016.

[36] Percona LLC. Percona TokuDB. https://www.percona.com/software/mysql-database/percona-tokudb, 2017.

[37] Lanyue Lu, Thanumalayan Sankaranarayana Pillai, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau. Wisckey: Separating keys from values in ssd-conscious storage. In *Proceedings of the 14th USENIX Conference on File and Storage Technologies (FAST 16)*, pages 133–148, 2016.

[38] Leonardo Marmol, Swaminathan Sundararaman, Nisha Talagala, and Raju Rangaswami. Nvmkv: a scalable, lightweight, ftl-aware key-value store. In *2015 USENIX Annual Technical Conference (USENIX ATC 15)*, pages 207–219, 2015.

[39] Neal Mielke, Todd Marquart, Ning Wu, Jeff Kessenich, Hanmant Belgal, Eric Schares, Falgun Trivedi, Evan Goodness, and Leland R. Nevill. Bit error rate in nand flash memories. In *Proceedings of the IEEE International Reliability Physics Symposium, (IRPS 08)*, pages 9–19. IEEE, 2008.

[40] MongoDB. MongoDB. https://www.mongodb.com, 2017.

[41] Dushyanth Narayanan, Eno Thereska, Austin Donnelly, Sameh Elnikety, and Antony Rowstron. Migrating server storage to ssds: analysis of trade-

offs. In *Proceedings of the 4th ACM European conference on Computer Systems (Eurosys 09)*, pages 145–158. ACM, 2009.

[42] Suman Nath and Aman Kansal. Flashdb: Dynamic self-tuning database for nand flash. In *Proceedings of the 6th International Conference on Information Processing in Sensor Networks*, pages 410–419. ACM, 2007.

[43] Netflix. Application data caching using ssds. `http://techblog.netflix.com/2016/05/application-data-caching-using-ssds.html`, May 2016.

[44] Patrick O'Neil, Edward Cheng, Dieter Gawlick, and Elizabeth O'Neil. The log-structured merge-tree (lsm-tree). *Acta Informatica*, 33(4):351–385, 1996.

[45] Oracle. Oracle Berkeley DB. `http://www.oracle.com/technetwork/database/database-technologies/berkeleydb/overview/index.html`, 2017.

[46] Pinterest. Open-sourcing rocksplicator, a real-time rocksdb data replicator. `http://bit.ly/2pv5nZZ`, 2016.

[47] William Pugh. Skip lists: A probabilistic alternative to balanced trees. *Algorithms and Data Structures*, pages 437–449, 1989.

[48] William Pugh. A Skip List Cookbook. Technical Report CS-TR-2286.1, University of Maryland, 1990.

[49] Pandian Raju, Rohan Kadekodi, Vijay Chidambaram, and Ittai Abraham. PebblesDB: Building Key-Value Stores Using Fragmented Log-Structured Merge Trees. In *Proceedings of the 26th Symposium on Operating Systems Principles*, SOSP '17, pages 497–514, Shanghai, China, October 2017. ACM.

[50] Parthasarathy Ranganathan. From microprocessors to nanostores: Rethinking data-centric systems. *Computer*, 44(1):39–48, 2011.

[51] Apache Samza. State management. `http://samza.apache.org/learn/documentation/0.8/container/state-management.html`, 2017.

[52] Russell Sears and Raghu Ramakrishnan. blsm: a general purpose log structured merge tree. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 217–228. ACM, 2012.

[53] Pradeep J. Shetty, Richard P. Spillane, Ravikant R. Malpani, Binesh Andrews, Justin Seyster, and Erez Zadok. Building workload-independent storage with vt-trees. In *Proceedings of the 11th USENIX Conference on File and Storage Technologies (FAST 13)*, pages 17–30, 2013.

[54] RocksDB Issue Tracker. Strategies to reduce write amplification 19. `https://github.com/facebook/rocksdb/issues/19`, 2014.

[55] Uber. Cherami: Uber engineering's durable and scalable queue in go. `https://eng.uber.com/cherami/`, 2016.

[56] Vijay Vasudevan, Michael Kaminsky, and David G. Andersen. Using vector interfaces to deliver millions of iops from a networked key-value storage server. In *Proceedings of the Third ACM Symposium on Cloud Computing (SOCC 12)*, page 8. ACM, 2012.

[57] Peng Wang, Guangyu Sun, Song Jiang, Jian Ouyang, Shiding Lin, Chen Zhang, and Jason Cong. An efficient design and implementation of lsm-tree based key-value store on open-channel ssd. In *Proceedings of the Ninth European Conference on Computer Systems (Eurosys 14)*, page 16. ACM, 2014.

[58] Xingbo Wu, Yuehai Xu, Zili Shao, and Song Jiang. Lsm-trie: An lsm-tree-based ultra-large key-value store for small data items. In *Proceedings of the 2015 USENIX Annual Technical Conference (USENIX ATC 15)*, pages 71–82, 2015.

[59] Demetrios Zeinalipour-Yazti, Song Lin, Vana Kalogeraki, Dimitrios Gunopulos, and Walid A. Najjar. Microhash: An efficient index structure for flash-based sensor devices. In *Proceedings of the 4th USENIX Conference on File and Storage Technologies (FAST '05)*, 2005.

[60] Huanchen Zhang, Hyeontaek Lim, Viktor Leis, David G. Andersen, Michael Kaminsky, Kimberly Keeton, and Andrew Pavlo. SuRF: Practical Range Query Filtering with Fast Succinct Tries. In *Proceedings of Special Interest Group on Management of Data*, SIGMOD '18, Houston, TX, USA, 2018. ACM.

Vita

Pandian Raju was born in Thoothukudi in the state of Tamil Nadu in India, the son of Mr. Raju Sankaran and Mrs. Shanthi Pandian. He did his schooling in Sakthi Vinayakar Hindu Vidyalaya in Thoothukudi. He secured the first rank in the entire state of Tamil Nadu in his 12^{th} board examination.

He received his Bachelor of Engineering degree in Computer Science from College of Engineering, Guindy (CEG), Anna University, Chennai. He did a summer internship (Software development role) at Amazon (in Bangalore) during his Bachelors. After his Bachelors, he worked as Software Development Engineer for two years in Flipkart (in Bangalore), which is one of India's biggest e-commerce companies. He was a part of the Data platform team in Flipkart, which mananges all the data in the company, where he helped create and maintain streaming data pipeline that powered the entire real-time analytics in the company.

He started his Masters in Computer Science at The University of Texas at Austin in Fall 2016, advised by Prof. Vijay Chidambaram. His masters research focuses on storage systems, while his areas of interests span across distributed systems, storage systems and key-value stores, big-data processing and management, blockchain technologies, and deep learning.

During his Masters, he did his summer internship (Software Engineer

role) at Quora, Mountain View, as part of the *distro* team, where he helped design and implement the re-architecture of email digest system. After Masters graduation, he is headed towards Rubrik, Palo Alto, where he will be working full-time as Software Engineer.

Permanent address: pandian4mail@gmail.com

This thesis was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.